CSC8631_C1011323_EDA-Report

Sahej Sareen

02/12/2021

Introduction

The demand for cyber security professionals has increased significantly in the last decade. This is because there has been substantial increase in the use of technology such as advance cell phone technology, high speed internet accessibility etc., which has led to increase in the data that is being generated. Thereby, the need to protect this data from vulnerabilities, risks, failures and attacks from hackers is imperative. The demand to build a cyber resilient strategy becomes a necessity to avoid reputation damages and protection from privacy invasion. There are now numerous programmes which prepare students to cope with the complex future of cyber security. Most of the courses related to cyber security are online and can be recorded or downloaded for further reference. These online courses access a network of world class academics and improve the courses handling. The Future Learn MOOC data set contains archetype-survey-responses, enrollments, leaving survey response, question response, statistical viewing of viewing, question response. The data recorded for the above observations are done in specific intervals.

In this project work the analysis is done on the statistical viewing of viewing data set. This data set contains varies parameters such as Title, total views, duration of the videos, location of videos where they were watched. Using the above data along with the guidance of Cross-Industry Standard Process for Data Mining (CRISP-DM) Methodology, exploratory data analysis is performed. The results obtained graphically and with the calculations helps gives an insight on the data which can then be used to solve the business problems.

Analysis

The above data analysis is done based on the CRISP-DM methodology. The following steps in this methodology are as follows a)Business Understanding b)Data Understanding c)data preparation d)modelling e)evaluation f)development

Q1) Cycle 1- How are the total views effected by the Step position.

In this question we need to understand whether there is any relation between the views and the step position. This is the business understanding step in the CRISP-DM methodology. The first step before beginning with our analysis, the data needs to be cleaned and the columns must be structured. Now comparing the columns of total views effected by the Step position in video_stats_3 which is for the month of September 2017. In order to find the relation between the total views and step position, the correlation function is applied. The value calculated is -0.7168596. This states that they are inversely proportionate to each other. Later the graph is plotted via ggplot() function. As per the graph we can conclude that the viewers decreases as they proceed through the course. Comparing the columns in the video_stats4 as well as for video_stats5 by repeating the same procedure. The results observed by calculation for videos statistics 4 is -0.673233 and for video stats5 is -0.6636169 . Using ggplot()for plotting the graph for video_stats 4 and video_stats 5, it can be inferred that the viewership starts to decrease as the course progresses towards the end.

Q2) Cycle 1 How many people who watched the videos have completely watched the videos till the end?

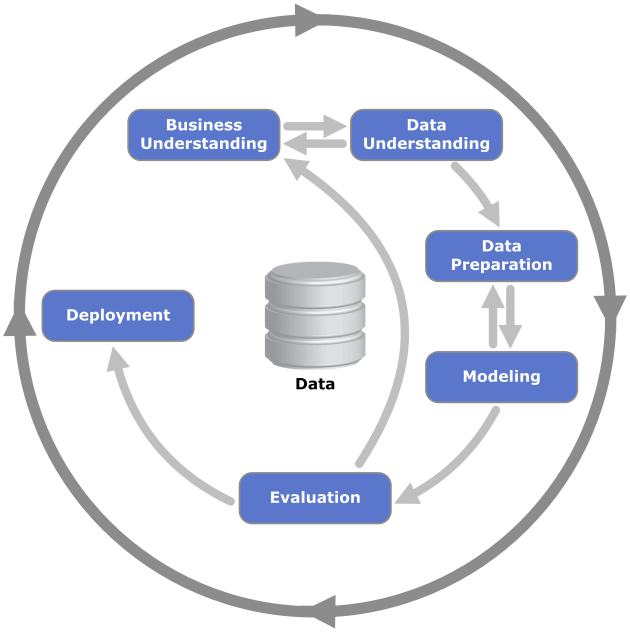

Figure 1: The CRISP-DM Model

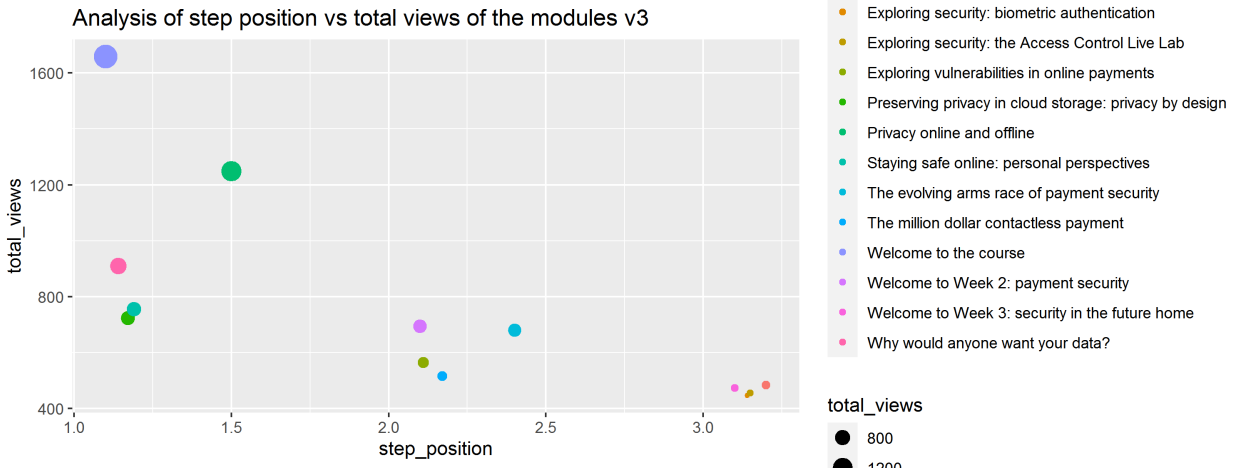

Figure 2: Analysis of step position vs total views of the modules v3

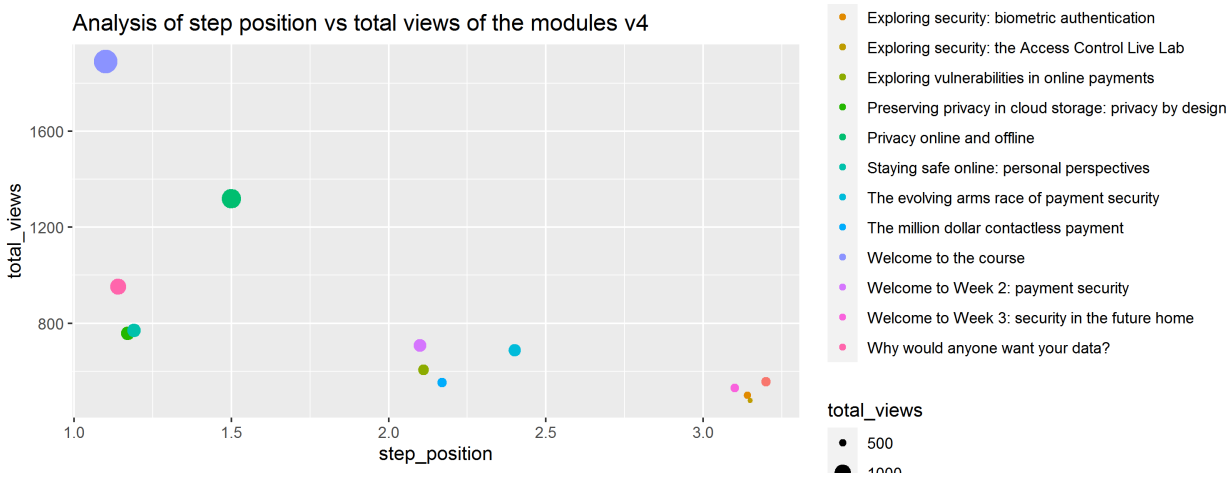

Figure 3: Analysis of step position vs total views of the modules v4

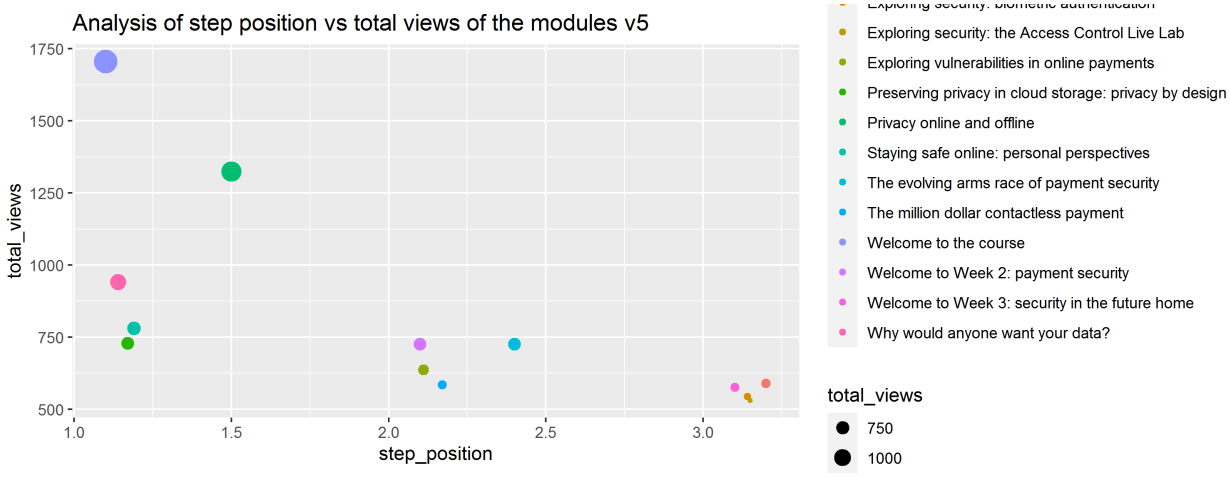

Figure 4: Analysis of step position vs total views of the modules v5

In this question we will be merging the data from the video_stats4 and video_stats5 i.e from November 2017 to February 2018. After combining the data files the correlation function is applied between columns total view and viewed onehundred percent to find the relation between them. From the calculated value the result obtained is 0.03268353. For plotting the graph the ggplot() function is used. As per the graph it can be stated that approximately 1850 has two values and the values above

From the correlation of the total views and onehundred percent it can be stated that: a) Whenever the total views increases not only the total learners watching 100 percent of video increases but also the percentage learners increases and vice versa. b) This shows that if any video has good content people viewing it increases at a healthy rate which is stated through positive numerical correlation observed here.

Cycle 2 What is the least and most common reason to leave the course? The data for this file is taken from the Leaving survey response. The data is transformed in data frame and the percentage of leaving reason variable is calculated. The data file for leaving response 4and5 are merged together. The data is then cleaned by omitting the blank columns. There are some text in the heading of the columns which needs to be changed to ensure that the readers can understand what is written. This ensure that there is no error while performing our analysis. The output is divided into seven categories. They are as follows a) I don't have enough time b) Other c) The course required more time than I realized d) The course was too easy e) The course was too difficult f) The course wasn't what I expected g) The course won't help me reach my goal